DOCUMENT RESUME

ED 431 958 CE 078 990

TITLE Measurement and Research Tools.
PUB DATE 1999-03-00
NOTE 32p.; In: Academy of Human Resource Development (AHRD) Conference Proceedings; see CE 078 963.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Administrators; Adult Education; Decision Making; *Educational Research; Employment Level; Factor Analysis; Foreign Countries; Human Resources; Influences; Job Performance; *Labor Force Development; *Measurement Techniques; Outcomes of Education; Participation; Professional Occupations; Questionnaires; *Research Methodology; Salary Wage Differentials; Test Construction; Test Items; Test Validity

ABSTRACT

This document contains four symposium papers on measurement and research tools. "Income Effects of Human Resource Development for Higher Educated Professionals" (Martin Mulder, Bob Witziers) reports on a study of 1,876 higher-educated professionals that found no correlation between participation in human resource development activities and subsequent job level or income. "Do Higher Performing Managers Actually Receive Better Ratings? A Validation of Multirater Assessment Methodology" (Allan H. Church), which is an applied example of a validation process based on a sample of 76 senior-level managers from a global health services organization, confirms the relationship between managerial performance and higher ratings on multirater feedback instrumentation. "Examination of the Tendency of Coefficient Alpha to Support the Choice of Instruments with Response Category Restriction" (Shani D. Carter) demonstrates that interpretation of coefficient alpha and the decision to use an instrument can be improved by considering item construction before deciding to use an instrument. "A Factor Analytic Investigation of the Content of Participant Reactions: Affective and Utility Judgments" (Ronald B. Morgan, Wendy Casper) presents the results of a study suggesting that participant reactions are multidimensional and "utility" judgments represent an underlying factor. (MN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 431 958

26 **Measurement and Research Tools**

Chair: Reid A. Bates, Louisiana State University

26 - 1 Income Effects of Human Resource Development for Higher Educated Professionals

Martin Mulder, University of Twente, The Netherlands

Bob Witziers, , University of Twente, The Netherlands

26 - 2 Do Higher Performing Managers Actually Receive Better Ratings? A Validation Of Multirater Assessment Methodology

Allan H. Church, W. Warner Burke and Associates, Inc.

26 - 3 Examination of the Tendency of Coefficent Alpha to Support the Choice of Instruments with Response Category Restriction

Shani D. Carter, University of Scranton

26 - 4 A Factor Analytic Investigation of the Content of Participant Reactions: Affective and Utility Judgments

Ronald B. Morgan, George Washington University

Wendy Casper, George Mason University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Mulder S Carter
A Church R Morgan

BEST COPY AVAILABLE

CE072590

Income Effects of Human Resource Development for Higher Educated Professionals

Martin Mulder
University of Twente

Bob Witziers
University of Twente

This paper addresses the question of whether individuals benefit from participating in HRD-activities. In particular the questoin is addressed whether participating in these activities lead to a higher job level and/or higher income. The results, obtained through means of a path analysis and based on a sample of 1876 respondents, show that participating in HRD-activities has no impact on individuals job level and income. The reverse is the case for other variables included in the analysis, such as gender, level of initial qualification and work experience.

Keywords: Human Resource Development, Adult Education, Outcomes

Introduction and Research Question

Research into the relationship between educational level and income has a long tradition. The initial steps were taken in the human capital theory developed by Becker (1975). In this theory students are thought to be investors, making a choice between spending time on income-generating activities and spending time on educational activities. In line with micro-economic thought, it is assumed that the income received as a result of time spent in education is the benefit the student strives to maximize in order to maximize his utility. The main focus of studies using this framework has been to determine the actual returns of education to the individual. In the Netherlands an important contribution is delivered by Glebbeek (1993), who has developed and studied a model for estimating effects of education on career development. He has found that education has a direct influence on income (.35), and an indirect influence via jobs persons held in the past (.71), jobs they hold later in their career (.45), and management functions they hold (.32). The magnitude of this indirect influence on income is .15. Peschar & Wesselingh (1995) conclude that this means that persons with the same jobs have different incomes according to the differences in their education. They state that this is according to the human capital theory that states that more education always pays.

The most recent development in this field are summarized by McMahon (1996). In this review he states that researchers over the last years have attempted to arrive at a comprehensive estimate of the current total returns to education as objective as possible, that is, to estimate these returns without overestimation or underestimation. As a result, economist have more than ever been able to detect the true effect of education on earnings. Moreover, these results show that some of the returns to education may have been seriously underestimated in the past.

Mincer (1979, 1989) has build on ideas formulated in the human capital theory by stating that individuals, after they have entered the labor market, keep investing in educational activities. Moreover, his theory predicts that these activities increase their human capital, which in turn leads to a higher income on the labor market. This notion is supported by several studies investigating the effects of participating in corporate training. For instance, a study conducted by Groot (1994) in the Netherlands showed that participants of corporate training earn 11% more than non-participating members of the work force. Similar results were found in studies conducted by Barron, Black & Loewenstein (1989), Holzer (1988) and Mincer (1989). However, other results can be found in studies conducted by Bonn (1993) and Tuynman (1989). These studies found that participation in HRD had no impact at all.

This paper builds on the notion developed by Mincer by exploring the effect of participating in HRD activities on the income of higher educated professionals. The relationship between HRD and income is difficult to study, as there are many variables related to the distribution of income. Work experience is one important example (Cohn & Geske, 1990), in particular since it can act as a substitute for training and education. Moreover, research has shown that there are gender differences and differences between workers in different economic sectors: ceteris paribus women end up in less managerial positions and earn less too, while workers in the profit sector earn more than workers in the non-profit sector (Meesters, 1993). Given these other factors affecting income, the research question we try to answer is:

Copyright © 1999 Martin Mulder and Bob Witziers

What is controlled for other relevant variables. the relationship between participation in human resource development and income of higher educated professionals?

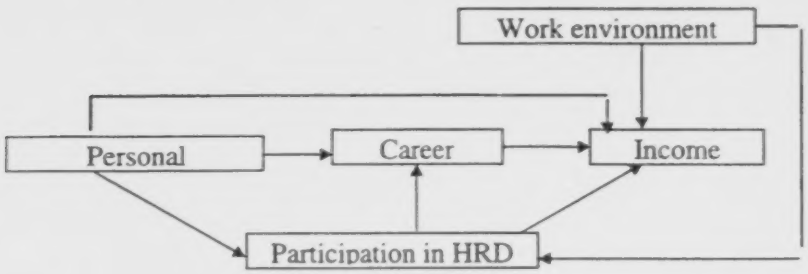
Methodology

Conceptual Framework and Operationalization

The data on which the study is based come from a study that is conducted for a national weekly magazine (with a readership of over 225,000) with a large number of advertisements for jobs for higher educated employees. This study started in 1995 with a base-line measurement. It was repeated with a slightly changed questionnaire in 1996 (Mulder & Witziers, 1998). And it will be repeated again in the years 1998 and 2000. The results presented in this paper are from the data that were collected in 1996.

The 16-page questionnaire comprises 63 questions with sub-questions in the following categories: 1. General questions; 2. Start on the labor market; 3. Career history; 4. Present situation; 5. Additional training; 6. Most important additional training program; 7. Quality of this additional training program; 8. Plans for further participation in training.

Based on the available data the following exploratory model was employed to study the research question at hand:



Firstly, this model proposes that individual characteristics (such as work experience, gender and initial educational level) affect individuals' career position, income and participation in HRD-activities. Secondly, it proposes that participation in HRD-activities affects directly individuals' career position and their income. Thirdly, it assumes that participation in HRD activities affects income indirectly by promoting their career position. Finally, the model takes into account that income and participation in HRD activities are a function of environmental factors such as the size of the organization and the economic sector of the organization. More specifically, the following variables were selected:

Personal characteristics

- gender (0 = female, 1 = male)
- number of years of work experience
- level of initial education (0 = vocational college; 1 = university);
- perceived quality of education (11 items);
- size of the job (number of hours an individual works per week);
- received a higher qualification of a regular educational institution (0 = no, 1 = yes)

Participation in HRD

- number of training programs taken per year
- amount of days participated in training programs per year

Career position

- job level (5 categories)

Work Environment

- size of the organization (number of employees);
- economic sector of the organization (0= profit; 1=non-profit).

Income

- gross earnings per annum

The meaning of most variables is clear, except (may be) the variable perceived quality of initial education. This variable measures the contribution of the initial education to finding a job. Therefore, participants in the survey were asked to denote whether or not their initial education contributed to aspects such as finding a job easily, finding a job with high salary, finding a job with good career prospects etc. The variable income (gross earnings per annum) is measured in seven blocks of 25.000 Dutch guilders (this equals with blocks of 11,261 Euro) each.

A path analysis was conducted to test the proposed model.

Sample

The questionnaire is published in the magazine on November 22, 1996. Apart from that 4,500 printed copies were sent by direct mail to a random selection from the readership. The number of respondents was 1,765. From the 4,500 questionnaires that were sent by direct mail, 859 came back; from the magazine, the remaining 906 were returned. Although the response rate is relatively small, the total number of respondents is large enough to make a valid statement about HRD for higher educated employees, and the relationship between HRD and income. Moreover, a non response analysis was carried out aimed at comparing the participation in HRD-activities of respondents with non respondents. No significant differences were found between the response group and the non-response group.

Results

Our analysis started with the model described in a previous section. However, our results show that, in order to get a better fit, some changes in the model had to be made. These changes concern opening the paths between the work environment factors and the career position variable.

Our final results (see table 1) show that the most important factors contributing to job level are gender, level of initial education, work experience and perceived quality of initial qualification. More specifically, this implies that higher positions are hold by men, university graduates, workers with more work experience and workers who's initial education gave them the opportunity to obtain a good job. Both indicators measuring the level of participation in HRD-activities have a negative relationship with job level. However, their contribution to this variable is not significant at the 0.05-level.

Table 1

Job level

Variable	Path coefficient	Standard error	T-Value
Gender	0.11	0.056	1.96
Level of initial qualification	0.31	0.043	7.20
Received higher diploma?	0.13	0.053	2.41
Work experience	0.05	0.004	12.57
Quality of initial education	0.13	0.032	4.22
Size of organization	- 0.13	0.015	- 8.67
Economic sector organization	- 0.18	0.050	- 3.56
Number of HRD-training programmes	- 0.03	0.019	- 1.56
Number of HRD-days	- 0.00	0.000	- 0.60

$R^2 = 0.18$

The results concerning the income variable show that all most all the variables included in the analysis have a significant relationship with income (see table 2). More specifically, this implies that the variables having a positive relationship with the variable job level, also affect directly the variable income. In other words, controled for job level, men have a higher income than female workers, university graduates earn more than vocational education graduates, workers in the profit sector earn more than workers in the non-profit sector. Moreover, workers in larger organizations and having more work experience earn more than workers in smaller organizations and workers having less work experience. Finally, respondents stating that their initial education gave them an easy and comfortable access to the labor market earn more than respondents stating this was less the case. Again, our results show that participation in HRD-activities does not have a significant contribution to income. As such, the main conclusion is that participation in HRD-activities has no (in)direct impact on respondents' income.

Table 2

Income

Variable	Path coefficient	Standard error	T-Value
Job level	0.45	0.023	19.85
Gender	0.32	0.054	6.00
Level of initial qualification	0.41	0.042	9.68
Received higher diploma?	- 0.11	- 0.051	- 2.19
Work experience	0.12	0.004	31.47
Quality of initial education	0.26	0.018	8.50
Size of job	0.22	0.024	9.03
Size of organization	0.12	0.014	8.35
Economic sector organization	- 0.32	0.048	- 6.68
Number of HRD-training programmes	- 0.03	0.018	- 1.43
Number of HRD-days	- 0.00	0.000	- 1.15

$R^2 = 0.62$

The results concerning participation in HRD-activities show that only a few variables have an impact the level of participation. The most important variable is work experience, indicating that in particular young people engage in HRD-activities.

Table 3

Number of HRD-Programmes

Variable	Path coefficient	Standard error	T-Value
Experience	- 0.06	0.004	- 15.48
Size of organization	0.08	0.018	4.69

$R^2 = 0.13$

Table 4

Number of HRD-Days

Variable	Path coefficient	Standard error	T-Value
Number of HRD-programmes	0.08	0.810	8.02
Received higher diploma	0.06	0.225	3.97
Work experiece	- 0.77	0.160	3.97
Economic sector	0.50	0.211	2.38

$R^2 = 0.07$

The results concerning model assessment show that our data fit the model reasonably well. Although the Chi-square test of exact fit suggests to reject the model given the small p-value, all other measures indicate that the data fit our model reasonably well.

Conclusion

The main conclusion of this paper is that most of the variables included in our model have a significant and positive relationship with job level and income. More specifically, gender, work experience, level of initial education, economic sector, size of job, perceived quality of initial education all have a significant relationship with either job level and/or income. Moreover, the direction of the relationships are in most cases in the expected direction. That is, our results are in line with either theoretical notions and/or results from previous research. However, it must be noted that variables relating to participating in HRD-activities do not have a relationship with the endogeneous variables job level and income. In other words, participants of HRD-activities do not benefit from their participation in terms of obtaining a higher position and/or getting a higher income. As such, these results are in line with other European studies (Boon, 1993; Tuynman, 1989). The results of these studies show that participating in HRD-activities does not contribute to a higher job level and /or a higher income.

Concerning the question of why participating in HRD does not pay off, in our view the most important reason has to do with the nature and context in which participation in HRD takes place. Most respondents state that they take part in HRD to develop themselves and, moreover, without informing their employee. In other words, the linkage between their HRD-activities and their job is weak. As such, the absence of a relationship between participation in HRD and benefits such as a higher job level and income is not a surprise. Moreover, respondents state that they participate in HRD to keep up with changes in their work situation. As such, this implies that their behavior is primarily oriented at continuing their present level than aimed at getting a better job. Therefore, it is no surprise that there is no relationship between participation in HRD and income.

Of course, our results can also be the result of the limitations of our study. For instance our study is limited with respect to the measurement of several variables, including the income variable and the measurement of participation in HRD. These deficiencies can mask the true effect of participating in HRD.

References

Barron, J., J. Bishop, J. & Loewenstein, M. (1989). Job matching and on the job training. Journal of Labor Economics, 7, 1-19.

Becker, G.S. (1975). Human Capital: A theoretical and empirical analysis with special reference to education.. NBER: Columbia University Press

Boon, M.J.J.P.M. (1993). Studeren bij de Open Universiteit als defensief onderwijsgedrag. Nijmegen: ITS.

Cohn, E. & Geske, T. (1990). The economics of education. Oxford Pergamon Press.

Glebbeek, A.C. (1988). De arbeidsmarkt van opleidingen. *Tijdschrift voor Onderwijsresearch, 4*, 75-89.

Glebbeek, A.C. (1993). *Perspectieven op loopbanen*. Assen: Van Gorcum.

Groot, W. (1994). Het rendement van bedrijfsopleidingen. Den Haag: Ministerie van bedrijfsopleidingen.

Holzer, H. (1988). The determinants of employee productivity and earnings. NBER Working Paper 2782.

McMahon, (1997). Recent advances in measuring the social and individual benefits of education. International Journal of Educational Research, 27(6), 453-482.
Mincer, J. (1974). *Schooling, experience, and earnings*. Washington: National Bureau of Economic Research.
Mincer, J. (1989). Job training: costs, returns and wage profiles. Columbia University.
Mulder, M. & Witziers, B. (1998). *Levenslang leren. Hoger opgeleiden en hun plannen*. Amsterdam: Intermediair.
Oosterbeek, H. (1992). *Essays on human capital theory*. Amsterdam: University of Amsterdam (dissertation).
Peschar, J. & Wesselingh, A. (1995). *Onderwijssociologie*. Groningen: Wolters-Noordhoff.
Tuynman, A. (1989). Recurrent education, earning and well-beings. Stockholm: Almqvist & Wicksell.

Do Higher Performing Managers *Actually* Receive Better Ratings? A Validation of Multirater Assessment Methodology

Allan H. Church
W. Warner Burke Associates Inc.

Despite significant cost and effort associated with feedback based OD/HRD interventions, relatively little attention has been paid to validation vis-à-vis hard performance measures. This study presents an applied example of a validation process based on a sample of 76 senior level managers from a global health services organization. Results confirmed the relationship between managerial performance and higher ratings on multirater feedback instrumentation, as well as its link with managerial self-awareness. Implications for practice are discussed.

Keywords: Multirater Feedback, Managerial Self-Awareness, Performance Ratings, Validation

In a society that is increasingly being driven by measurement and metrics (Church & Waclawski, 1998a; Hronec, 1993; Swanson, 1994), it should be no surprise that the process of assessing and developing managerial behavior has become one of the core practice areas of applied organizational scientists. While more traditional single-source methods of performance assessment are still common enough in organizations (Bretz, Milkovich, & Read, 1992; Murphy & Cleveland, 1995; Schuler & Huber, 1990), more and more HRD, OD and I/O practitioners are turning to development and appraisal applications based on data collected from multiple sources (e.g., Antonioni, 1996; Bracken, 1994; Church, 1995a; O'Reilly, 1994; Shipper & Neck, 1990). Though the terms vary considerably (e.g., multirater, multisource, full circle, 360-degree, and upward feedback), the interest in this process as a performance and change management tool is evident.

Based on a combination of action research and survey feedback (e.g., Lewin, 1951; Nadler, 1977), one of the fundamental assumptions of the multirater approach is that data about one's specific workplace behaviors, when compared with internal (self) perceptions, can be used as a catalyst for enhancing self-awareness and subsequent behavioral change (Church & Waclawski, 1998b; London, 1997; McLean, 1997; Tornow, 1993; Yammarino & Atwater, 1997). Fortunately for all those involved in implementing such systems, research results supporting the link between individual performance and managerial self-awareness have begun to reach critical mass (e.g., Atwater & Yammarino, 1992; Church, 1997; 1994; Church & Waclawski, 1999; Furnham & Stringfield, 1994; Van Velsor, Taylor, & Leslie, 1993). Moreover, while there are still significant questions regarding the efficacy of feedback itself as a tool (Kulger & DeNisi, 1996), some research does support its impact on behavior change when used in conjunction with OD/HRD interventions (Cheston, 1996; Hazucha, Hezlett, & Schneider, 1993; London & Wohlers, 1991).

Thus, in spite of the myriad of issues and concerns associated with implementing multirater feedback systems—e.g., regarding confidentiality, validity, rater training, cultural support, etc—current estimates of usage run anywhere from 12% to 29% (Mercer's Fax Facts..., 1996; Waldman, Atwater, & Antonioni, 1998) and appear to be growing still. Moreover, when one considers the extent to which multirater methods have already been employed in executive and management development efforts in the past (e.g., Bernardin, 1986; Burke, Richley, & DeAngelis, 1985; Church, Javitch & Burke, 1995; Hazucha et al., 1993; London & Beatty, 1993; Nowack, 1992), the total use percentage is likely to be even higher. In short, the notion of using data from a variety of perspectives for developing and improving individual effectiveness has moved beyond being just another consulting fad—it is now a legitimate OD, HRD and I/O intervention methodology.

Given the establishment and indeed institutionalization of multirater feedback methods in some organizations, however, practitioners have begun to question whether these technologies may be being adopted too quickly and recommend a host a cautions and warnings regarding implementation (e.g., Antonioni, 1997; Church & Bracken, 1997; Church & Waclawski, 1998b; McLean, 1997; Waldman et al., 1998). At this point, what is needed is a more concerted effort at developing a meaningful body of theory and research to effectively guide practice.

One critical issue in the area of multirater research and practice, that has received relatively little attention, is the nature of the multirater instrumentation used itself (Church & Waclawski 1998b, McLean, 1997). It is a common (though probably not often articulated) assumption among many feedback practitioners, for example, that

Copyright © 1999 Allan H. Church

higher ratings on a given multirater assessment instrument--almost regardless of its purported content--will directly translate to or correlate with some degree of improved individual performance. This is one of the central reasons why a large number of practitioners and consulting firms feel comfortable specializing in developing customized instrumentation for multirater interventions based on key issues in the client organization (Church & Waclawski, 1998b). The idea is, with enough focused attention and one-on-one coaching and interpretation, almost any set of practices will provide some level of meaningful information for the focal manager. Oddly enough, this is also the same rationale for justifying the use of standardized behavioral measures across different (and arguably unique) organizational settings. Unfortunately, however, these assumptions, particularly in the former situation, are rarely tested. Instead, it is common practice, especially among non-research based practitioners and firms, for feedback systems to be designed, implemented and then ultimately abandoned before the notion of data-based validation has ever been truly considered. Although there are a myriad of reasons for this situation--e.g., cost, effort and issues of confidentiality associated with linking individual feedback to existing internal performance measures, not to mention the fear of finding out that perhaps the multirater instrument constructed does not in fact correctly differentiate among levels of performance—there is arguably an ethical responsibility on the part of feedback practitioners to validate these types of applied instruments. Clearly, regardless of the amount of time, money, or effort being directed at a given OD or HRD feedback effort, if the instrument being used for assessment purposes lacks validity and reliability from a measurement perspective (Nunnally, 1978), or applicability for the organization from a practice perspective (Church & Waclawski, 1998b; Waldman et al., 1998), it will not matter what the data collection process reveals about an individual's developmental needs.

For the most part, however, the issue of instrument validation has belonged to the realm of I/O Psychology related selection and appraisal applications (e.g., Murphy & Cleveland, 1995), rather than OD and HRD based developmental programs. In fact, this may be one of the contributing factors to the "softer" connotations often associated with these latter types of interventions (Church & Burke, 1995). In any case, despite the importance of this seemingly obvious measurement issue, few organizations take the necessary steps to validate the use of an instrument for their own employees, regardless whether it is a standard or customized measure. In other words, although many internal and external OD, HRD and I/O practitioners work with feedback processes on the belief that a measurement instrument derived from set of core values, practices, or competencies (i.e., whatever the source) will indeed differentiate among more and less effective performers, relatively few validations of this assumption are ever tested empirically.

The purpose of the following study, then, is to contribute to the feedback literature by describing the results of an applied case example of how this multirater assessment tool validation process can be done using existing data from various feedback and personnel related resources. A secondary interest is to attempt to replicate prior findings (e.g., Church, 1997) regarding the link between managerial self-awareness and performance. More specifically, the following research questions were explored:

- Do higher performing managers receive better (i.e., more positive) ratings from their co-workers compared to lower performers?
- Are higher performing managers more self-aware (i.e., exhibit greater self-other ratings congruence) of their own behaviors?

In short, this paper will detail both a process and a set of results for exploring the validation of a multirater instrument. The findings will be relevant to both practitioners *and* researchers working with feedback related systems.

Method

Sample

Data for the present study were collected from 76 senior level managers (and 308 direct reports, 291 peers and 71 supervisors of the same individuals) from a global health services organization. These individuals represented a subset (for which additional differentiating performance evaluations were made available) of a larger group (n=389) of attendees of one of 12 senior management training and development programs conducted by an external consulting firm during a two year period. Participation in the program was based primarily on position in the organization and scheduling availability. In terms of overall attendee demographics, 78% of the participants were male and 22% female. Regarding tenure, 8% indicated having recently joined the organization, 20% had been there

2-5 years, 22% had 6-10 years tenure, and the remaining 50% had been with the organization for more than 11 years. In terms of age, approximately 20% were under 39 years old, while the majority (45%) were in their forties. Although geographic location was distributed among areas including Africa, Asia, Australia and New Zealand, Canada, Central/South America, Europe, Japan, and the UK, the majority of respondents (59%) were based in the U.S. with the next largest group (18%) being stationed in Western Europe.

Procedure

As pre-work for the program, all participants were required to complete a self-report behavior instrument and return it directly to an independent external firm for report purposes. In addition, parallel forms of the behavioral questionnaires were distributed to the direct reports, peers and supervisors of these same individuals for completion and subsequent return. Before completing the questionnaires, all raters, including the participants themselves, were familiarized with the developmental intent of the rating process and the manner in which the feedback would be used. Raters were instructed to be as candid and honest in their assessments as possible. All the data obtained remained anonymous throughout the feedback process.

Additional data were provided by the HRD/OD function of the organization for a subset of these attendees in order conduct a validation analysis—i.e., to link responses obtained from the multirater assessment process to actual performance evaluations. Selection into the subgroup for this study was based on the extent to which individuals were considered to be either *consistently above* or *consistently below* average in their overall performance over a two year period using archival formal appraisal ratings from personnel records as the criterion. All performance evaluations had been collected from supervisors substantially prior to selection to or attendance in the development program. Thus, this performance measure was entirely independent of the multirater feedback process at the direct report and peer level.

Although a larger number of individuals (n=100) were initially identified based on prior performance, some managers had not yet attended one of the senior management development seminars at the time of the study and others had been to a program offered earlier with different data collection procedures and thus were excluded. The final group for analysis consisted of 76 executives in total—i.e., 41 higher performers (rated by 170 direct reports, 162 peers and 39 supervisors), and 35 lower performers (rated by 138 direct reports, 129 peers and 32 supervisors).

Measures

The behavioral assessment tool used was a custom designed instrument that had been developed for this organization based a new set of core competencies and strategic leadership objectives. The instrument contained 61 items pertaining to various aspects of managerial behavior. For each item respondents were asked to rate the focal individual using a scale from 1 to 7 (where 1 = seldom, 3 = sometimes, 5 = often, and 7 = almost always) on the degree to which that senior manager engaged in a given behavior (e.g., " gives others clear feedback regarding their performance, " cooperates with others"). Content areas included developing and rewarding employees, promoting individual and organizational learning, resolving conflict, clearly communicating to others, building relationships within and across functions, focusing on strategic and business issues, and making decisions. Overall, the internal reliability of the instrument was quite high across each rating perspective ranging from .95 for self to.98 for supervisors.

A second "outcome" measure of work group climate (Burke, 1986) was also distributed to focal managers and their direct reports. Used in prior feedback programs and research (e.g., Church, 1995b; Church & Waclawski, 1997; Litwin & Stringer, 1968; Litwin, Humphrey & Wilson, 1978), this 28 item measure assessed perceptions of the climate experienced in the work group on a day-to-day basis along six dimensions: clarity, recognition, participation, standards, intraunit relations and interunit relations. In order to be consistent with the metrics of the behavioral measure, a 7 point scale was also employed. In general, the internal reliabilities for the summary scores for the instrument were also quite high among the present sample (.94 and .97).

Analyses

The data were analyzed at both the summary score and item levels using a variety of techniques including MANOVA, and simple oneway ANOVAs and t-tests. Averaged other ratings by source were used where multiple ratings existed (e.g., direct report and peer). Managerial self-awareness vis-à-vis self-direct report congruence was computed using two different methods: (a) the collective level of agreement between self and direct report ratings using a standard index of profile similarity *d*, i.e., the square root of the sum of squared item differences between

self and averaged direct reports ratings (e.g., Church, 1997, 1994; Nunnally, 1978); and (b) the four category agreement method (e.g., Yammarino & Atwater, 1997) whereby participants are classified as being either under-raters, *lower behaving* accurate-raters, *higher behaving* accurate raters, or over-raters based on a comparison of self and direct report total scores. Standard correlations were also performed for comparison purposes with prior research.

Results & Discussion

In general, the data were examined for a variety of performance and source (i.e., self, direct report, peer, supervisor) related effects across several different levels of analysis. Results are described below according to two major categories.

Multirater Feedback and Performance

First, a simple analysis of the overall summary scores by ratings source revealed the typical results—i.e., weak or nonexistent correlations—among self and others' ratings (see Table 1 for details). Consistent with prior ratings studies (e.g., Church, 1997; Harris & Schaubroeck, 1988; Nowack, 1992; Van Velsor et al., 1993), these senior managers tended to over-rate themselves compared with their direct reports ($t = 2.12$, $p < .05$). It was interesting to note, however, that peer ratings did not significantly differ in level from self-assessments, and supervisors provided the highest set of ratings overall across all 61 behaviors compared with direct reports ($t = 4.67$, $p < .001$) and peers ($t = 2.74$, $p < .01$).

Table 1.Descriptive Statistics for Summary Ratings

			Behaviors				*Climate*	
	M	*SD*	*Self*	*Drs.*	*Prs.*	*Supv.*	*Self*	*Drs.*
Behavior								
Self	5.51	.62	.95					
Direct Reports	5.27	.62	.05	.98				
Peers	5.48	.47	.21	.37*	.97			
Supervisors	5.70	.75	.04	.40*	.41*	.98		
Climate								
Self	5.40	.73	.75*	.24	.24	.25*	.94	
Direct Reports	5.03	.66	.09	.79	.36*	.43*	.37*	.97

Note: * $p < .05$, reliabilities are printed on the diagonal

These latter findings are more consistent with prior research on influence or laterally-based positions (e.g., Church & Waclawski, 1999) suggesting that the management of peer and supervisor relationships were particularly important for these executives. Although self perceptions of climate were also significantly higher than direct report ratings on this measure ($t = 3.80$, $p < .001$), the level of these ratings for both groups overall were significantly lower than they were for the behavioral assessments (self, $t = 2.37$, $p < .05$, and direct report $t = 4.60$, $p < .001$).

Next, the primary objective of this research was addressed—i.e., the relationship between managerial performance and multisource ratings—using an MANOVA model to look for overall effects across summary scores by performance group. Since the overall omnibus F was significant $F(4,56) = 5.96$, $p < .001$, additional univariate models were examined. As anticipated, the outcome of the univariate models (see Table 2) confirmed that higher performing managers were indeed rated significantly more positively by their co-workers (direct reports and peers) across the behavioral measure compared with their lower performing counterparts. In addition, direct reports of higher performing managers indicated experiencing significantly more positive climates in their work groups compared with those of lower performers.

Table 2. Univariate ANOVA Results for Ratings Source by Performance

		Lower Performers		*Higher Performers*	
	F	*M*	*SD*	*M*	*SD*
Behavior					
Self	0.01	5.51	.67	5.50	.57
Direct Reports	3.81*	5.11	.60	5.39	.61
Peers	3.90*	5.36	.53	5.58	.38
Supervisors	30.01**	5.25	.79	6.07	.47
Climate					
Self	2.41	5.26	.76	5.53	.68
Direct Reports	6.92**	4.81	.67	5.22	.61

Note: * $p < .05$, ** $p < .01$.

When examined at the item level for direct report ratings, 11 behaviors were found to significantly differentiate between higher and lower performers. In general, these centered on four major content areas: building relationships, decision making, problem solving/conflict resolution, and providing feedback & motivating employees. A similar set of content areas emerged when the peer data was examined at the item level as well. More specifically, 18 different peer assessments (in the areas of learning & skill development: problem solving/conflict resolution, integration with the larger organization, customer focus, and communication) served to discriminate among executives in the two performance groups. With respect to climate, 13 of the 28 descriptors yielded significant differences among work groups as well suggesting greater levels of clarity in roles and goals, standards of performance for the group, positive and friendly relations among members, and relationships and coordination with other workgroups in the organization. Although only a subset of behaviors and climate areas were identified, primarily having to do with relationships within and between functions, it is important to note that the trend in means was extremely consistent across almost all items and probably would have reached significance had the sample size been larger. In any case, these findings appear to confirm the notion that higher performing executives do indeed exhibit effective managerial behaviors more frequently and foster stronger or more positive work group climates.

Supervisor ratings also conformed to this relationship, yielding a particularly strong degree of differentiation among performance. In fact, when examined at the item level it was apparent that almost all (i.e., 58 out of 61) of the behaviors assessed served to significantly differentiate between higher and lower performers. While this should not be surprising given the non-independence of the initial performance appraisal measure (i.e., supervisors were the only group that would have been previously aware of the performance of the focal executives), nonetheless, this high degree of consistency does further support the validity of the multirater assessment process. Further, given the already present degree of inflation among supervisors ratings, the confirmation of the central relationship here should be encouraging for practitioners. In other words, even with major rating biases and level effects a multirater feedback tool can be strong enough to correctly differentiate among higher and lower behaving managers and, thus by implication, their performance as well.

Aside from the confirmation of this central relationship, one interesting and somewhat surprising finding here was the total absence of significant differences among self-ratings by performance for either behavior or climate. Although one might expect better performers to rate themselves higher on each measure as well (e.g., whether due to knowledge of past performance, greater levels of self-awareness, or simply inflated egos), this was not the case. Instead, both groups of executives were equally likely to over-rate themselves compared with observations from direct reports and peers. Perhaps this reflects a modesty bias on the part of higher performers, a trend which has been suggested in the past (Church, 1997; Yammarino & Awtaer, 1997). Alternatively, given their senior positions in the organization, this trend may reflect a relatively consistent positive ego perception on the part of this sample. One way to explore this trend further, of course, is through an analysis of ratings congruence.

Managerial Self-Awareness and Performance

Next, the data were examined for differences with respect to managerial self-awareness. Although prior research (e.g., Church, 1997; Church & Waclawski, 1999) has shown an empirical relationship between self-awareness and performance, the differences observed here between self and direct report behavior ratings, though sizeable, did not reach statistical significance. The same test performed on climate ratings, however, did support the

notion that higher performing managers were more aware of the strengths and weaknesses in their work unit's climate than lower performers (see Table 3).

Table 3. ANOVA Results for Profile Similarity by Performance

		Lower Performers		*Higher Performers*	
	F	*M*	*SD*	*M*	*SD*
Behavior: Self-Direct Reports	1.82	1.36	.42	1.23	.36
Climate: Self-Direct Reports	5.80*	1.38	.52	1.13	.28

Note: * $p < .05$. Lower profile similarity values indicate greater item-by-item congruence.

When these same data were examined using the four group categorical approach (e.g., Yammarino & Atwater, 1997), which is intended to avoid the problems associated with difference scores, however, the outcome did support the link between performance and managerial self-awareness at the behavioral level (Chi-square of 7.60, $p < .05$). Although the number of individuals in each cell were low, nonetheless, a significantly greater proportion of the higher performers comprised the accurate higher behaving group (64%) than the accurate lower behaving one (25%). In addition and as suggested earlier, while there were no differences by performance in the over-rating group, modesty may also been a factor in that a much larger proportion of higher performers compared to lower performers (71% vs. 28%) comprised the under-rating group. Interestingly enough, although the trend in percentages was very similar, the differences in group composition for the climate ratings did reach significance using this method of self-other comparison.

Although one could argue based on these different outcomes that methodology plays a role here, given prior research, it is more likely that the power to detect significant effects in the present sample of executives was simply too limited. Future studies over time and with larger samples should be able to help us determine the definitive answer to this question. For now, however, these results do continue, at least in part, to support the relationship between self-other ratings congruence (i.e., managerial self-awareness) and hard measures of managerial performance.

Limitations & Suggestions for Future Research

In general, there are several limitations that limit the generalizability of these results. Most significant is the nature of the sample itself. Inclusion in the present study was limited to a small set of very senior level executives from a single health services organization, all of whom had also attended a feedback-based developmental program. Although this is clearly not an ideal sample in terms of size or diversity, it is a common enough example of the realities of working with and linking existing datasets and measures midstream in an organizational setting. Unfortunately, using what is available in an organizational system often results in unavoidable issues with respect to sample size (and therefore reduced power to detect effects) as well as a variety of potential selection issues. The fact that any effects were present at all given the size of the present sample should provide some indication of the importance of this relationship. In any case, future research should focus on replicating the effects demonstrated here across a variety of samples from different levels, functions within and across various organizations and industries. Another limitation in this research is the lack of other types of hard performance measures. While supervisor appraisal records over a two year period due represent an independent and arguably "hard" measure of managerial performance (after all, this is the information by which organizations make personnel decisions), the addition of other types of indicators such as sales figures, departmental turnover ratios, size of budgets, span of control, etc. could have helped to substantiate the effects reported even further. Whenever possible, researchers need to work hard to gather and incorporate this information into their analyses.

References

Antonioni, D. (1996). Designing an effective 360-degree appraisal feedback process. *Organizational Dynamics, 25*(2), 24-38.

Atwater, L. E., & Yammarino, F. J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology, 45*, 141-164.

Bernardin, H. J. (1986). Subordinate appraisal: A valuable source of information about managers. *Human Resource Management, 25*, 421-439.

Bracken, D. W. (1994). Straight talk about multirater feedback. *Training & Development, 48*(9), 44-51.

Bretz, R. D., Milkovich, G. T., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications, *Journal of Management, 18*, 321-352.

Burke, W. W. (1986). *Executive practices and climate questionnaire.* Pelham, NY: W. Warner Burke Associates

Burke, W. W., Richley, E. A., & DeAngelis, L. (1985). Changing leadership and planning processes at the Lewis Research Center, National Aeronautics and Space Administration. *Human Resource Management, 24*, 81-90.

Cheston, S. L. (1996). *Managerial self-awareness: A longitudinal study*. Unpublished Doctoral Dissertation, Graduate School of Arts and Sciences, Columbia University, New York.

Church, A. H. (1997). Managerial self-awareness in high performing individuals in organizations. *Journal of Applied Psychology, 82*(2), 281-292.

Church, A. H. (1995a). First-rate multirater feedback. *Training & Development, 49*(8), 42-43.

Church, A. H. (1995b). Managerial behaviors and work group climate as predictors of employee outcomes. *Human Resource Development Quarterly, 6*(2), 173-205.

Church, A. H. (1994). Managerial self-awareness in high performing individuals in organizations. *Dissertation Abstracts International, 55-05B*, 2028. (University Microfilms No. AAI9427924).

Church, A. H., & Bracken, D. W. (1997). Advancing the state of the art of 360-degree feedback: Special issue editors' comments on the research and practice of multi-rater assessment methods. *Group & Organization Management, 22*(2), 149-161.

Church, A. H. & Burke, W. W. (1995). Practitioner attitudes about the field of organization development. In W. A. Pasmore and R. W. Woodman (Eds.), *Research in organizational change and development* (Vol. 8, pp. 1-46). Greenwich CT: JAI Press.

Church, A. H., & Javitch, M., & Burke W. W. (1995). Enhancing professional service quality: Feedback is the way to go. *Managing Service Quality, 5*(3), 29-33.

Church, A. H., & Waclawski, J. (in press). How to make friends and influence people: An investigation of influence behaviors and effectiveness in lateral relations. *Human Resource Development Quarterly, 10*(1).

Church, A. H., & Waclawski, J. (1998a). *Designing and using organizational surveys*. Aldershot, England: Gower.

Church, A. H., & Waclawski, J. (1998b). Making multirater feedback systems work. *Quality Progress, 31*(4), 81-89.

Church, A. H., & Waclawski, J. (1997). *The impact of leadership style on work group climate and enablement.* Paper presented at the 4th Annual International Conference on Advances in Management, July 10, Rexdale, Ontario Canada.

Furnham, A., & Stringfield, P. (1994). Congruence of self and subordinate ratings of managerial practices as a correlate of supervisor evaluation. *Journal of Occupational and Organizational Psychology, 67*, 57-67.

Harris, M. and Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*, 43-61.

Hazucha, J. F., Hezlett, S. A., & Schneider, R. J. (1993). The impact of 360-degree feedback on management skills development. *Human Resource Management, 32*(2-3), 325-351.

Hronec, S. M. (1993). *Vital signs: Using quality, time, and cost performance measurements to chart your company's future*. New York, NY: AMACOM.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

Lewin, K. (1951). *Field theory in social science*. New York: Henry Holt.

Litwin, G. H., & Stringer, R. A. (1968). *Motivation and organizational climate*. Boston: Harvard Business School Press.

Litwin, G. H., Humphrey, J. W., & Wilson, T. B. (1978). Organizational climate: A proven tool for improving performance. In W. W. Burke (Ed.), *The cutting edge: Current theory and practice in organization development* (pp. 187-205). La Jolla, CA: University Associates.

London, M. (1997). *Job Feedback: Giving, seeking and using feedback for performance improvement.* Mahwah, NJ: Lawrence Erlbaum.

London, M., & Beatty, R. W. (1993). 360-degree feedback as a competitive advantage. *Human Resource Management, 32*(2&3), 353-372.

London, M., & Wohlers, A. J. (1991). Agreement between subordinate and self-ratings in upward feedback. *Personnel Psychology, 44*, 375-390.

McLean, G. N. (1997). Multirater 360 feedback. In L. J. Bassi & D. Russ-Eft (Eds.) *What works: Assessment, development, and measurement* (pp. 87-108). Alexandria, VA: American Society for Training and Development.

Mercer's fax facts survey results: Multisource assessment (November 20, 1996). William M. Mercer, Inc.

Murphy, K. R & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA; Sage.

Nadler, D. A. (1977). *Feedback and organization development: Using data-based methods.* Reading, MA.: Addison-Wesley.

Nowack, K. M. (1992). Self-assessment and rater-assessment as a dimension of management development. *Human Resource Development Quarterly, 3,* 141-155.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

O'Reilly, B. (1994). 360° feedback can change your life. *Fortune, 130*(8), 93-94, 96, 100

Schuler, R. S., & Huber, V. L. (1990). *Personal and human resource management* (4th ed.). St. Paul, MN: West Publishing Company.

Shipper, F., & Neck, C. P. (1990). Subordinates' observations: Feedback for management development. *Human Resource Development Quarterly, 1,* 371-385.

Swanson, R. (1994). *Analysis for improving performance. Tools for diagnosing organizations & documenting workplace expertise.* San Francisco: Berrett-Koehler.

Tornow, W. W. (1993). Editor's note: Introduction to special issue of 360-degree feedback. *Human Resource Management, 32*(2-3), 211-219.

Van Velsor, E., Taylor, S. & Leslie, J. B. (1993). An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness. *Human Resource Management, 32*(2&3), 249-263.

Waldman, D. A., Atwater, L.E., & Antonioni, D. (1998) Has 360 degree feedback gone amok? *Academy of Management Executive, 12*(20) 86-94

Yammarino, F. J., & Atwater, L. E. (1997). Do managers see themselves as others see them? Implications of self-other rating agreement for human resources management. *Organizational Dynamics, 25*(4), 35-44.

Examination of The Tendency of Coefficient Alpha to Support The Choice of Instruments With Response Category Restriction.

Shani D. Carter
University of Scranton

This paper examines the effect of item construction on coefficient alpha and the effect of coefficient alpha on the comparison of instruments measuring similar constructs. It is predicted that decreasing the number of response categories of instrument items increases alpha for an instrument, but the response category restriction also decreases construct validity. Interpretation of alpha and the decision to use an instrument can be improved by considering item construction before deciding to use an instrument.

Keywords: Coefficient Alpha, Evaluation, Construct Validity

Previous research has found that job-related outcomes such as training effectiveness, post–training behavior, job satisfaction, and turnover, are significantly influenced by personal characteristics, such as motivation, affect (i.e., mood), and self-efficacy (Endler, 1976; Goldstein, 1993; Judge, 1992; Kane, 1986; Kanfer & Ackerman, 1989; Noe, 1986; Pervin, 1978; Terborg, 1981; Wexley, 1984). These personal characteristics influence performance even when ability and environmental factors remain constant, and therefore should be measured as part of any study of training outcomes (i.e., reaction, knowledge, behavior, and results; Kirkpatrick, 1959).

Measurement of Personal Characteristics

Instrument design

Personal characteristics such as affect and motivation are generally measured through self-report questionnaires using Likert scales (DeVellis, 1991). These questionnaires are based on theory, and are used when it is difficult to measure the construct in another way (DeVellis, 1991; Nunnally, 1978). These questionnaires generally contain 5 to 30 questions, where respondents are required to indicate their degree of agreement or disagreement with a statement. The degree of agreement or disagreement is generally restricted to a three-point scale (e.g., no agreement, neutral, agreement), a five-point scale, or a seven-point scale. Question responses are usually summed, so that the possible range of scores of an instrument is directly related to the range of possible responses to individual questions.

Numbers of Response Categories

Instruments with a higher number of response categories per question are able to measure a greater level of detail of the range of differences between respondents. For example, for a questionnaire with 10 questions, if there are three response categories per question, the range of scores for the questionnaire is 0 to 30, but if there are five response categories per question, the range of scores for the questionnaire is 0 to 50. The questionnaire using five response categories therefore reports the range of responses between participants in much greater detail than the questionnaire with only three response categories. This greater detail of measurement of personal characteristics can lead to higher construct validity (i.e., more accurate portrayal and measurement of the construct) which can lead to greater criterion-related validity (i.e., prediction of job-related outcomes).

Examining Reliability and Validity

Effect of Reliability and Validity on Generalizability

Using reliable and valid instruments to measure personal characteristics leads to greater generalizability of conclusions about job-related outcomes (Carmine & Zeller, 1979; Cook & Campbell, 1979; DeVellis, 1991; Nunnally, 1978). If a researcher uses a poorly constructed instrument, the researcher may be led to the erroneous conclusion that a particular personal characteristic had no effect on job-related outcomes, although the lack of

Copyright © 1999 Shani D. Carter

correlation was actually due to the instrument's poor construction. It is therefore crucial to ensure that instruments used to measure personal characteristics are of the highest reliability and validity possible.

Measuring Reliability

Researchers commonly use coefficient alpha to examine the reliability of instruments before deciding whether to include instrument scores as variables in studies of job-related outcomes, such as training evaluation analyses (Cronbach, 1970; Peterson, 1994). Some researchers rely exclusively on coefficient alpha as the indicator of the reliability and usefulness of an instrument. This heavy reliance on alpha can lead a researcher to design an instrument that can measure only small amounts of variability, because coefficient alpha is based upon the inter-item correlations of an instrument. For example, a researcher may choose to use an instrument with items that are answered on a three point scale rather than a five point scale because the three point scale allows for less possible variability of scores and thus produces higher inter-item correlations and higher alpha.

Drawbacks of Exclusive Reliance on Alpha

Previous researchers have discussed the drawbacks of relying on coefficient alpha as an indicator of the reliability and usefulness of an instrument without also examining other characteristics of the instrument such as actual correlations and instrument length (e.g., Cortina, 1993; Barchard & Hakistian, 1997; Schmitt, 1996; Thompson, 1990; Zimmerman, et al., 1993). For example, Cortina (1993) and Schmitt (1996) recommended examining the effect of the number of questions of an instrument, and also examining the dimensionality of the instrument before deciding whether to trust coefficient alpha as an accurate gauge of an instrument's reliability. Barchard and Hakistian (1997) recommended examining confidence intervals when using coefficient alpha. Thompson (1990) recommended examining the statistical relationship of individual items to the instrument as a whole when interpreting alpha. Finally, Zimmerman, et al. (1993) recommended examining the correlations of subtest scores when interpreting alpha.

Hypotheses

The current paper extends this research by suggesting that item construction also should be examined before determining whether to trust coefficient alpha and whether to include an instrument in research on job-related outcomes such as training evaluation. *It is predicted that decreasing the number of response categories of instrument items will increase the instrument's mean inter-item correlation, thereby increasing alpha for the instrument.*

The response category restriction decreases construct validity which decreases criterion-related validity. Exclusive use of alpha for analysis of reliability of an instrument can then lead a researcher to the erroneous decision to use an instrument with weak construct validity and criterion-related validity, which can then lead to the erroneous conclusion that a construct has no significant effect on job-related outcomes.

Method

Subjects

A repeated measure design was used. Two sessions were conducted where subjects completed questionnaires during weeks 1 and 5. Of the 93 subjects enrolled in the study, ten subjects did not complete the two sessions, and so were dropped from further analysis.

Measures

The current research uses Affective Trait as the construct to be examined. Affective trait is the tendency to experience a general emotional state (i.e., mood), and can be either positive or negative (George, 1991). Although negative affect does not influence either recall or learning, positive affect does influence recall of material (Isen, Daubman, & Nowicki, 1987). When in a state of positive affect, individuals recall more material and more diverse material, which leads to more creative and efficient problem solving (Isen, et al., 1985). Because affect has been

shown to influence learning, it is a useful variable to be included in training evaluation, and the results of this study can be used to design instruments for future training evaluation research.

Affective trait was measured with the Positive and Negative Affect Schedule (PANAS), which measures positive affect and negative affect as two independent dimensions (Watson, Clark, & Tellegen, 1988). The measure contained 20 items, where positive affect (PA) and negative affect (NA) were measured with 10 items each. Responses were on a 5-point scale, where 1 was low and 5 was high. Responses were summed, so that the range of possible scores for the positive affect and negative affect sections was 10 to 50 each.

Affective trait also was measured with the Neutral Objects Satisfaction Questionnaire, (NOSQ; Judge, 1992). This instrument measures affective disposition as an affective bias toward items that are common in everyday life. The instrument contains 25 items, and has three response categories, where 1 is dissatisfied, 2 is neutral, and 3 is satisfied. This instrument will be used as a comparison for the examination of the effects of response category restriction of the PANAS.

Analyses

The following analyses were conducted to examine the effects of response category restriction on the PANAS.

1. Correlations and factor analyses of the PANAS and NOSQ were conducted to determine the extent to which they measure the same underlying construct.
2. Participant responses to PANAS items were recoded from a scale of 1 to 5 to a scale of 1 to 3. This recoding simulated PANAS scores if the items had been constructed with 3 response categories. The specific method of the recoding was done as follows: Actual responses of 1 or 2 were recoded 1; actual responses of 3 were recoded 2; and actual responses 4 or 5 were recoded 3. Thus, actual responses in the low range (1 or 2) were still coded low (1), actual responses that were neutral (3) were still coded neutral (2), and actual responses that were high (4 or 5) were still coded high (3).
3. Coefficient alpha was computed and compared for the PANAS actual and recoded scores.
4. Correlations and factor analyses were then repeated using the PANAS recoded scores and NOSQ actual scores.

Results

PANAS and NOSQ Variable Correlations

Correlation matrices were produced for the NOSQ and PANAS variables. These contained correlations for session 1 and session 2 for: the actual and recoded PA variables; the actual and recoded NA variables; and the actual NOSQ variables. The mean inter-item correlations and alpha statistics derived from these tables are summarized in Table 1. Alpha was computed as follows: (alpha = (k*r)/(1+(k-1)*r), where k = the number of items and r = the mean inter-item correlation). Table 1 show that there were small differences between the mean inter-item correlations for the actual and recoded PA and NA variables. For PA, the mean inter-item correlation increased from .26 to .27 with recoding in session 1, but decreased from .52 to .49 with recoding in session 2. For NA, the mean inter-item correlation decreased from .30 to .26 with recoding in session 1, and decreased from .36 to .30 with recoding in session 2.

Table 1

Mean Correlations and Alpha Statistics for PANAS and NOSQ

	Correlation Session 1	Correlation Session 2	Alpha Session 1	Alpha Session 2
PA	0.26	0.52	0.78	0.91
PA recode	0.27	0.49	0.79	0.91
NA	0.30	0.36	0.81	0.85
NA recode	0.26	0.30	0.78	0.81
NOSQ	0.11	0.15	0.76	0.81

PANAS and NOSQ Coefficient Alpha

Alpha statistics for the above discussed correlations also are presented in Table 1. For PA, alpha increased from .78 to .79 with recoding in session 1, but remained at .91 with recoding in session 2. For NA, the alpha decreased from .81 to .78 with recoding in session 1, and decreased from .85 to .81 with recoding in session 2. Alpha for the NOSQ was .76 at session 1 and .81 at session 2.

Factor Analysis of NOSQ and PANAS Variables.

Results of factor analysis using varimax rotation for the actual and recoded PA and NA variables are presented in Table 2. This table shows that the PANAS measured two constructs (PA and NA), and that there were few significant differences in the magnitude of the factor loadings between the actual and recoded PA and NA variables. In addition, the variance explained by each factor remained largely consistent between the actual and recoded variables.

Factor analysis of the NOSQ was unsuccessful due to nonconvergence. Therefore, the high alphas for the NOSQ for sessions 1 and 2 (.76 and .81) are used as indicators of the instruments unidimensionality. Factor analysis of the NOSQ and PANAS variables together also was unsuccessful due to nonconvergence.

Analysis of NOSQ and PANAS Constructs

The NOSQ, PA, and NA variables were summed to produced the respective constructs. Table 3 contains correlations for the NOSQ and PANAS constructs. This table shows that the correlations for the actual and recoded PA and NA variables were nearly 1.0. The correlation between the actual and recoded PA was .97 for session 1 and .98 for session 2. The correlation between the actual and recoded NA was .96 for session 1 and .93 for session 2. There were differences in the correlations between the NOSQ and the PA and NA constructs for the actual and recoded PA and NA variables. For session 1, the correlation between NOSQ and PA increased from .31 to .35 with recoding and from .19 to .24 with recoding for session 2. For session 1, the correlation between NOSQ and NA decreased from -.23 to -.19 with recoding and from -.22 to -.19 with recoding for session 2.

Factor analysis of the NOSQ and PANAS constructs was again unsuccessful due to nonconvergence.

Table 2

Factor Analyses of Actual and Recoded PANAS Variables

	Actual PANAS		Recoded PANAS	
	Factor 1	Factor 2	Factor 1	Factor 2
PA1	-0.01	**0.73**	0.03	**0.76**
PA2	0.05	**0.70**	-0.02	**0.65**
PA3	0.03	**0.60**	0.01	**0.60**
PA4	0.20	**0.46**	0.11	**0.53**
PA5	0.00	**0.57**	-0.07	**0.46**
PA6	-0.27	**0.48**	-0.31	**0.43**
PA7	-0.21	**0.49**	-0.23	**0.56**
PA8	-0.34	**0.27**	-0.31	**0.35**
PA9	-0.32	**0.18**	-0.35	**0.26**
PA10	-0.31	**0.46**	-0.33	**0.41**
NA1	**0.67**	0.10	**0.65**	0.03
NA2	**0.70**	-0.04	**0.73**	-0.11
NA3	**0.57**	0.10	**0.55**	0.13
NA4	**0.53**	0.07	**0.49**	-0.06
NA5	**0.55**	0.03	**0.49**	0.02
NA6	**0.65**	0.13	**0.60**	0.02
NA7	**0.40**	-0.18	**0.33**	-0.17
NA8	**0.44**	-0.17	**0.36**	-0.24
NA9	**0.60**	-0.22	**0.59**	-0.18
NA10	**0.32**	-0.22	**0.26**	-0.20
Var. exp. by factor	3.57	2.90	3.28	2.91
Final communality. est. Total		6.48		6.19

Table 3

Correlations of NOSQ and PANAS Constructs

		1	2	3	4	5	6	7	8	9	10
1	Sess. 1 PA	1.00									
2	Sess. 1 PA recode	0.97	1.00								
3	Sess. 2 PA	0.61	0.62	1.00							
4	Sess. 2 PA recode	0.63	0.65	0.98	1.00						
5	Sess. 1 NA	-0.20	-0.24	-0.15	-0.17	1.00					
6	Sess. 1 NA recode	-0.27	-0.32	-0.20	-0.22	0.96	1.00				
7	Sess. 2 NA	-0.16	-0.18	0.08	0.04	0.55	0.47	1.00			
8	Sess. 2 NA recode	-0.18	-0.20	-0.02	-0.07	0.55	0.50	0.93	1.00		
9	Sess. 1 NOSQ	0.31	0.35	0.17	0.20	-0.23	-0.19	-0.26	-0.23	1.00	
10	Sess. 2 NOSQ	0.29	0.32	0.19	0.24	-0.14	-0.14	-0.22	-0.19	0.81	1.00

Discussion

The hypothesis that decreasing the number of response categories of instrument items would increase alpha for the instrument was not supported with this sample. Although there were small differences between correlations for the actual and recoded variables and constructs, these differences were inconsistent for PA and NA. This result could be due to the fact that few respondents used the extreme ends of the PANAS scale (i.e., 1 or 5), such that the recoding had only a small effect on the PA and NA variables.

Nevertheless, the assumption that decreasing the number of response categories of instrument items would alter the correlation between the instrument and another instrument was supported. Specifically, the correlation between PA and NOSQ increased with recoding of PA for session 1 and session 2. Further, the correlation between NA and NOSQ decreased with recoding of NA for session 1 and session 2. These results indicate that research could inadvertently overestimate or underestimate the construct validity or criterion-related validity of an instrument if the instrument has response category restriction.

Conclusions

Although the hypothesis regarding alpha was not supported, the results indicated that response category restriction does affect construct validity and criterion-related validity when an instrument is validated against another instrument. It is therefore recommended that instruments used to measure personal characteristics should contain 10 to 20 items, and that items should be constructed with scales containing five to seven response categories. Instruments with low numbers of items should contain more response categories, and vice versa. An instrument constructed in this way can measure a sufficient level of differences between respondents, but is not so long or tedious as to fatigue the respondents. In order to determine whether to use an instrument after data is collected, researchers should examine: (1) the coefficient alpha; (2) the dimensionality of the instrument through factor analysis; (3) inter-item correlations; and (4) the relationship of the instrument to a similar construct. The results of these four statistical analyses should be interpreted in light of the number of items and number of response categories of the instrument.

References

Barchard, K. A. & Hakstian, A. R. (1997). The effects of sampling model on inference with coefficient alpha. *Educational and Psychological Measurement, 57,* 893-905.

Carmine, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-001). Beverly Hills, CA: Sage.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston: Houghton Mifflin.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

DeVellis, R. F. (1991). *Scale development: Theory and applications* (Sage University Paper Series on Applied Social Research Methods, 26). Beverly Hills, CA: Sage.

George, J. M. (1991). State or trait: effects of positive mood on prosocial behaviors at work. *Journal of Applied Psychology, 76,* 299-307.

Goldstein, I. L. (1993). *Training in organizations: needs assessment, development, and evaluation.* Montery, California: Brooks/Cole Publishing.

Isen, A. M., Daubman, K. A., & Nowicki, G. P. (1987). Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology, 52,* 1122-1131.

Isen, A. M., Johnson, M. M. S., Mertz, E., & Robinson, G. F. (1985). The influence of positive affect on the unusualness of word association. *Journal of Personality and Social Psychology, 48,* 1-14.

Judge, T. A. (1992). The dispositional perspective in human resources research. *Research in Personnel and Human Resources Management. 10,* 31-72

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: an integrative/aptitude-treatment approach to skill acquisition. *Journal of Applied Psychology, 74,* 657-690.

Kirkpatrick, D. L. (1959). Techniques for evaluation training programs. *Journal of the American Society of Training Directors, 13,* 3-9.

Murray, H. A. (1938). *Explorations in personality.* New York: Oxford University Press.

Noe, R.A. (1986). Trainee attributes and attitudes: Neglected influences on training effectiveness. *Academy of Management Review, 4,* 736-749.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill

Pervin, L. A. (1978). Theoretical approaches to the analysis of individual-environment interaction. In L. A. Pervin & M. Lewis (Eds.), *Perspectives in interactional psychology* (pp.67-86). New York: Plenum Press.

Peterson, R. A. (1994). A meta-analysis of Cronbach's Coefficient Alpha. *Journal of Consumer Research, 21,* 381-391.

Silver, N. C. & Dunlap, W. P. (1987). Averaging correlation coefficients: Should fisher's z transformation be used? *Journal of Applied Psychology, 72,* 146-148.

Terborg, J. R. (1981). Interactional psychology and research on human behavior in organizations. *Academy of Management Journal, 6,* 569- 576.

Thompson, B. (1990). Alphamax: A program that maximizes coefficient alpha by selective item deletion. *Educational and Psychological Measurement, 50,* 585-589.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales, *Journal of Personality and Social Psychology, 54,* 1063-1070.

Wexley, K. N. (1984). Personnel training, *Annual Review of Psychology, 35,* 519-551.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53,* 33-48.

A Factor Analytic Investigation of the Content of Participant Reactions: Affective and Utility Judgments

Ronald B. Morgan
Wendy Casper
George Mason University

The premise of this research is that increased understanding of the content of participant reactions is a necessary step toward improving the construct validity of those judgments. This study examines the factor structure of a large data base of participant reactions, and explores the construct validity of "utility" judgments as defined by Alliger et.al. Results suggest that reactions are multidimensional, and that "utility" judgments represent an underlying factor.

Keywords: Training Evaluation, Participant Reactions

Participant reactions are by most accounts the principal means by which organizations evaluate training programs (Bassi, Benson & Cheney, 1996; Saari, Johnson, Mclaughlin & Zimmerle, 1988). According to Training Magazine (1996) approximately 86 percent of organizations use participant reactions to evaluate their training efforts. Furthermore, for most of those organizations, participant reactions are the sole means of evaluation. Consequently, improving the technology of participant reaction will yield great benefits. The purpose of this research is to identify strategies for improving participant reactions as a training evaluation methodology.

Kirkpatrick (1959; 1996) describes four levels of training evaluation. This framework has served as the organizing scheme for both practitioners and researchers. Level I represents the reactions of participants to training, typically collected through questionnaires administered at the conclusion of the training program. The second level of evaluation consists of measures of actual learning in the training program. This typically involves post training measures of knowledge, skill or attitude change. In essence, it represents an examination of the "internal validity" of the training effort. Level III evaluation is focused on gauging the extent to which learning is "transferred" to the job, and typically consists of traditional measures of job performance including ratings provided by superiors, peers or subordinates. Finally, Level IV evaluation assesses changes in organization outcomes (revenue, profit, turnover, etc) attributable to the training program.

Because Level II, III and IV outcome data are far more costly to plan and collect, organizations tend to hold their training program to the minimal level of outcome, that is, to determine simply whether or not participants react favorably to the program (Bassi et al., 1996).

The Uses of Participant Reactions

To understand the potential for an improved technology of participant reaction, it is helpful to discuss the various uses reactions serve. First, participant reactions can provide substantive input for the design and improvement of training efforts. Brinkerhoff (1986) describes a continuous improvement process, with participant reactions providing one source of data for continuous improvement of each component of the training effort. Presumably, credible reaction data can be a tremendous diagnostic tool for improving the design and delivery of training. Second, the process of collecting participant reactions can serve a "public relations" function, building up the customer's sense that the training function is sincerely interested in the service they deliver. A number of authors have argued for the importance of "customer orientation" and the measurement of customer satisfaction in human resources (Heneman, Huett, Lavingna and Ogsten, 1995).

A third use for participant reactions is as a surrogate for other, more costly criteria of training effectiveness- measures of learning, measures of on the job behavior or performance, and measures of organization results. Because of the efficiency of the methodology (e.g., surveys collected at the conclusion of training) the development of reactions as predictors of actual learning, transfer to the job, and organizational results, would dramatically increase the utility of training evaluation.

Copyright © 1999 Ronald B. Morgan & Wendy Casper

The Relationship of Trainee Reactions and Other Evaluation Criteria

A number of studies have examined the correlation among the four levels of training criteria, with generally discouraging results (Alliger, et. al, 1997; Alliger and Janek, 1989). The meta-analytic results presented in those studies show little systematic relationship between traditional participant reactions and other levels of training criteria. In fact, the low magnitude of correlation of trainee reactions and other criteria have led some (Holman, 1996) to question the appropriateness of participant reactions as a criterion of training effectiveness.

While trainee reactions in general show little systematic relation to learning, job performance and organization results, particular facets or dimensions of trainee reactions appear to hold great promise. Alliger et. al (1997) and Alliger and Janek (1989) distinguish between "affective" and "utility" judgments. Affective judgments measure the extent to which the participant "liked" or was satisfied with different components of the training. Utility judgments, on the other hand, reflect the extent to which the participant can apply the content to their job. These researchers present evidence to suggest that "utility" judgments show modest but significant relationship to immediate learning (level II) (r = .26) and to transfer (level III) (r = .18). Furthermore, studies reporting combined scale of affective and utility judgments show significant relationship to immediate learning (level II) (r = .14), to behavior (skill) learning, a level II distinction made by those researchers (r = .12), and to on the job use (level III) (r = .21).

Understanding the Underlying Constructs of Participant Reactions

While the Alliger et al. (1997) research is promising in that it illustrates the potential utility of specific judgments, a number of observations must be made. First, the distinctions drawn by the research to date (Alliger and Janek, 1989; Warr and Bunce, 1995; Alliger et al., 1997) have characterized clusters of items which appear on evaluation forms, but have not presented any factor analyses of any multi-item reaction forms. While these constructs may well be relevant and important characterizations of trainee reactions, there has been no empirical basis to suggest that trainee reactions are adequately summarized by either Alliger et al. framework of "affective" and "utility", or Warr and Bunce's (1995) "affective," "utility" and "difficulty" dimensions. Alliger and Janek (1997) present reliability estimates for the affective and utility judgments, and evidence that the item clusters relate to other criteria, they do not present evidence to suggest that those two dimensions adequately account for the variability of participant reactions. While this research offers great promise for increasing the effectiveness of training evaluation, it must be noted that the dimensions they identify are conceptually based. No factor analysis of comprehensive participant reactions have been reported. Clearly, continued development of the technology of training evaluation will benefit from identification of the underlying dimensionality of participant reactions.

It is the premise of this research that an understanding of the underlying constructs of participant reactions will improve their predictive utility. In effect, the current researchers assert that the low magnitude of relationships between participant reactions and other criteria (Alliger & Janak, 1989; Alliger, et al., 1997) are due to an inadequate understanding of the diverse content of participant reaction instruments. An increased understanding of the "facets" of participant reactions will allow researchers and practitioners to achieve stronger predictor - criterion relationships. This premise is based in part on the evolution of the job satisfaction literature. An extensive body of literature investigated the relationship between job satisfaction and performance, and repeatedly found low relationships ranging from .14 - .17 (Brayfield & Crockett, 1955; Iaffaldano & Muchinsky, 1985) leading many researchers to conclude little relationship existed between employee satisfaction and employee job performance. However, other researchers (Fisher, 1980) proposed that the low relationship between satisfaction and performance may be due to measuring satisfaction and performance at different levels of specificity, such as measuring general job satisfaction but then measuring performance on a very specific task.

More recently, this same argument has been made to explain the low relationships typically found between personality and job performance. Since the early 1990s (Barrick & Mount, 1991), there has been a plethora of research on the value of personality in predicting performance (Barrick, Mount, & Strauss, 1993; Hogan, Hogan, & Murtha, 992; Hough, 1992; Ones, Mount, Barrick, & Hunter, 1994; Tett, Jackson, Rothstein, & Reddon, 1994). Research has typically found the percent of variance personality predicts in job performance to be small. However, recently Borman, Hansen and Hedge (1997) suggested that these low relationships might be due to using supervisory ratings as the performance criterion. They argued that supervisory ratings typically reflect both task and contextual aspects of performance, but that if a more specific measure of contextual performance was used, the personality-performance relationship may be larger than previous research has suggested (Borman, Hanson, & Hedge, 1997). One study provides support for this notion in finding that supervisor ratings typically do represent both task and contextual aspects of performance (MacKenzie, Shankster, & Kohler, 1991).

Thus, this same argument could be made to explain the low relationships between trainee reactions and other measures of training evaluation when over-general composite constructs have been used (Alliger & Janak, 1989) and the increased magnitude of relationships when more focused constructs such as "utility" have been used (Alliger et al., 1997). However, the current researchers assert that participant judgments will exhibit greater complexity than the constructs (utility, reaction and difficulty) discussed to date.

The Content of Participant Reactions

A fundamental principle of measurement is that the construction of a measure requires a clear understanding of the construct to be measured. Understanding the construct is a necessary step toward improving any measure. This is true of the measurement of participant reactions as well. It is our suggestion that research on participant reactions will benefit from an exploration of the entire content domain of participant reactions. To date, there has been little research that explored the full domain of participant reactions.

A number of opinions may be relevant as a means to capture participant reaction to training programs. Furthermore, the various facets of participant reactions may well predict different criteria. Phillips (1997) identifies 15 content areas (Progress with objectives, Program content, Instructional materials, Pre-work materials, Assignments, Method of delivery, Instructor/Facilitator, Motivation to learn, Program relevance, Registration/logistics, Facilities, Overall evaluation, Planned improvements, Recommendations for target audiences, and Instructor). Each of these content areas can be further broken down into a variety of relevant issues. For example, Phillips breaks the "Instructor" category into seven issues (preparation for sessions, knowledge of subject matter, presentation skills, communication skills, responding to learners, use of appropriate technology, and encouraging application) (Phillips, p. 115-116).

Clearly, the content domains of participant reactions are numerous and each domain likely consists of various issues and perceptions. Furthermore, each of these content domains may well serve different purposes in a systematic training evaluation effort. For example, certain dimensions may well be most useful in predicting customer satisfaction; others in predicting learning; others (utility judgments) in predicting use on the job, and still others in predicting organizational results attributable to training.

Judgments

In addition to measuring diverse content areas, a variety of judgments can be asked from participants. While no taxonomy of judgments is available in the literature, one can imagine measures of amount (too much - too little), measures of adequacy or effectiveness (judged against the individuals standards or against some defined or objective standards) and measures of satisfaction (judged against the participants needs or expectations). The diversity of judgments adds to the complexity of characterizing the dimensionality of participant reactions.

The Current Research

The current research examines the dimensionality of comprehensive participant reactions. This research is based on the premise that a refined understanding of the dimensions or facets of participant reactions will increase the predictive validity and utility of those judgments.

The Dimensionality of Participant Reactions

Three research questions emerge from the literature on participant reactions. First, what is the underlying factor structure of comprehensive participant reactions? It is reasonable to predict that the typical participant reaction form, organized into content areas (Phillips, 1998) will consist of content based dimensions. That is, items dealing with the instructor will co-vary; items dealing with the material will co-vary, and so forth. Furthermore, it is reasonable to suggest that these individual facets of participant reaction will have relevance to specific criteria. That is, judgments about course content will predict specific criteria, reaction data about instructor delivery will predict other criteria, etc. The central point is that only by mapping individual predictor--criterion relationships will the psychometric properties of participation reactions be improved. The second question to address is whether a more parsimonious set of factors can explain variability in participant reactions. Alliger et al. (1997) suggest that much of the diverse content of reaction forms collapses into a single "affective" dimension. Their model summarizes all of the variability in participant reactions into two dimension, "affective" and "utility" judgments. A

third question to address is whether or not empirical evidence emerges for the conceptual distinctions drawn by Alliger et al. In particular, this research will examine whether "utility" judgments emerge as an underlying factor in participant reactions. Finally, in addition to exploring the emergence of a "utility" factor, this research will investigate the amount of variance accounted for by these dimensions, and by doing so, explore their relative impact on overall trainee reactions.

Method

Procedures

Participant reaction forms were collected from 9,128 employees of a 750,000 employee organization. The form was used to collect level 1 evaluations for a variety of non-managerial training programs conducted over a three year period. Participants were asked to report some basic demographic data as well as their reactions to the training program. Participant responses were anonymous, as there was no identifying information (i.e. name, social security number) reported on the form. Data collected from participant reactions was included in a large data base which was later used to conduct the analysis reported in this paper.

Participants

About half (48.8%) were male, 36.1% were female and 15.3% did not report their sex. More than half of the respondents were White (54%), 19.7% were Black, 3.1% were Asian, 4.2% were Hispanic, 1.1% were Native American, 17.1% of the respondents did not report their race. Most of the employees were 40-49 years old (38.9%) or 30-39 years old (27.4%), while 12% were 50-59 years old, 1.2% were 60 or older, 2.5% were 17-29 and 18% did not report their age.

Items and scales

The reaction form consists of 32 items, organized into 8 categories. All 32 items ask the respondent to rate the satisfaction with aspects of training.

Results

Principle Axis Factoring with Oblimin rotation was performed on the 32 items which made up the participant reaction form (see Table 2), with the squared multiple correlation used to estimate communalities. The Kaiser criterion was applied to the initial solution, suggesting that six factors with Eigenvalues greater than one should be extracted. In addition, examination of the scree plot suggested the presence of six underlying factors among the items. Examination of the initial extraction shows that the first six factors accounted for a total of 62.8% of the variance. Eigenvalues and percent of variance accounted for by each of the factors are presented in Table 1.

After the initial extraction, the pattern matrix was rotated to examine for more interpretable loadings. We chose to rotate and examine the pattern matrix instead of the structure matrix because in using the oblique rotation method we were interested in the unique variance accounted for by each factor. Since the pattern matrix yields partial weights, the values in this matrix are more appropriate to interpret. In order to guide examination of weights in the pattern matrix, a .45 cutoff for the loading, which represented 20% of the variance, was required to retain the item with the factor. Finally, all the items which loaded on a given factor were examined in an attempt to discern what latent construct was underlying each factor. Table 2 presents the 32 items categorized into the 6 rotated factors, and the means and standard deviations for items. Table 3 presents the factor loadings

Table 1
Eigenvalues and percent variance accounted for by each factor

Factor	Eigenvalue	Percent of Variance
1	11.07	34.6
2	3.51	11.0
3	1.73	5.4
4	1.39	4.3
5	1.29	4.0
6	1.12	3.5

Factor 1 appears to represent a "satisfaction with instructor" dimension. Items 12, 13, 14, 15, 16, and 17 loaded on this factor. Factor 2 appears to represent "overall satisfaction with the training process". Items 26, 27, 28, 29, 30, 31, and 32 load on this factor. Factor 3 appears to represent a satisfaction with the testing process, with items 23, 24 and 25 load on this factor. Factor 4 represents a "utility of training" factor. Items 2, 3, 6, 8, 20, 21 and 22 load on this factor. In some ways, this factor may mirror what Alliger, et al (1997) referred to as utility reactions. Factor 5 represents a "materials" factor. Items 4, 5, 18, and 19 load on this factor. Factor 6 appears to represent a "course structure" factor, with items 9, 10, and 11 loading on this factor.

To examine the intercorrelation of the constructs, factor derived scale scores were created. The intercorrelation of the six constructs are presented in Table 4.

Table 2
Table of participant reaction item, associated factors, and item descriptive statistics

	Item	Factor	M	SD
1	How satisfied are you with communication of course objectives in clear, understandable terms?	-	4.5	.64
2	How satisfied are you with match of course objectives with your idea of what would be taught?	4	4.31	.75
3	How satisfied are you with the relevance of the course content to your job?	4	4.43	.71
4	How satisfied are you with the quality of course materials?	5	4.41	.74
5	How satisfied are you with the audio and visual aids used by the instructor?	5	4.33	.75
6	How satisfied are you with course's emphasis on most important info.?	4	4.39	.72
7	How satisfied are you with the order the course material was presented?	-	4.39	.68
8	How satisfied are you with realism of course presentation and class exercises?	4	4.30	.78
9	How satisfied are you with the length of training course?	6	2.42	.86
10	How satisfied are you with the pace of the course material presented?	6	2.75	.56
11	How satisfied are you with the amount of student participation in the course?	-	2.82	.55
12	How satisfied are you with the instructor's knowledge of course material and subject matter?	1	4.70	.54
13	How satisfied are you with the instructor's ability to keep interest of class?	1	4.53	.72
14	How satisfied are you with the instructor's presentation and explanation of course materials?	1	4.55	.67
15	How satisfied are you with the instructor's responsiveness to student questions and problems?	1	4.65	.60
16	How satisfied are you with instructor ability to relate to students individually?	1	4.62	.62
17	How satisfied are you with the instructor's overall effectiveness?	1	4.60	.66
18	How satisfied are you with the supplies and equipment for this course?	5	4.34	.77
19	How satisfied are you with classrooms, furniture, learning environment, etc.?	5	4.28	.86
20	How satisfied are you with the extent to which the course prepared you to perform current job tasks more effectively?	4	4.19	.78

Table 2 (continued)

21	How satisfied are you with the extent to which the course prepared you to perform new job tasks not currently performed?	4	4.14	.81
22	How satisfied are you with the quality of this training course overall?	4	4.42	.73
23	How satisfied are you with the fairness of the course exam?	3	4.16	.77
24	How satisfied are you with exam coverage and importance of material tested?	3	4.16	.77
25	How satisfied are you with feedback you received as result of course testing?	3	4.15	.80
26	How satisfied are you with the availability of training courses for individuals in your job classification?	2	3.58	1.11
27	How satisfied are you with the communication of training information to employees in your facility?	2	3.46	1.14
28	How satisfied are you with the quality of training services provided locally?	2	3.67	1.05
29	How satisfied are you with quality of training services provided regionally?	2	3.63	1.01
30	How satisfied are you with registration process and information you received prior to arrival at training?	2	3.71	1.09
31	How satisfied are you with quality of training courses provided by management?	2	3.86	.90
32	How satisfied are you with relevance of training you received for specific job functions?	2	3.84	1.09

Table 3
Table of participant reaction items and factor loadings

Item #	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Item 1				--		
Item 2				-.5566		
Item 3				-.5834		
Item 4					.5448	
Item 5					.4495	
Item 6				-.4666		
Item 7				--		
Item 8				-.4013		
Item 9						.4592
Item 10						.7153
Item 11						--
Item 12	.7304					
Item 13	.7322					
Item 14	.7276					
Item 15	.8105					
Item 16	.8071					
Item 17	.8274					
Item 18					.7976	
Item 19					.5147	
Item 20				-.8017		
Item 21				-.7148		
Item 22				-.5492		
Item 23			-.9393			
Item 24	-		-.9584			
Item 25			-.8292			
Item 26		.7261				

Item 27	.7940
Item 28	.7633
Item 29	.8102
Item 30	.4930
Item 31	.6641
Item 32	.6054

Table 4
Correlations between factor-based scale scores derived from factor analysis

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
[1]Factor 1	1.000					
[2]Factor 2	.2179	1.000				
[3]Factor 3	.4260	.3168	1.000			
[4]Factor 4	.7141	.3618	.5233	1.000		
[5]Factor 5	.5753	.2666	.4125	.6460	1.000	
[6]Factor 6	.2513	.1499	.1333	.2692	.2008	1.000
Mean	4.61	3.67	4.15	4.31	4.37	2.58
Stand Dev	.55	.82	.74	.61	.66	.59

[1]instructor satisfaction factor
[2]overall training satisfaction
[3]satisfaction with testing
[4]utility of training
[5]materials satisfaction
[6]course structure satisfaction

An interesting pattern of interrelationship exists among the factors, with the utility factor strongly related to "satisfaction with instructor" (r = .71), satisfaction with course testing (r = 52), and satisfaction with course materials (r = .65). The intercorrelation of the factor derived scales is presented in Table 4.

Conclusion and Discussion

The first, and most compelling finding from this research is the emergence of the "utility" factor described by Alliger et al. (1997). All four of the items presumed to measure the "usefulness" of training load on the same dimension, even though the items appear on different parts of the reaction questionnaire. Because Alliger et al. (1997) have demonstrated that utility judgments hold potential as predictors of learning and subsequent on the job use of the training content, this "empirical" support for the construct validity of the "utility" dimension is noteworthy.

A second finding of interest is that the instructor dimension (Factor 1) appears to carry great weight in trainees' overall perceptions of the training effort. The instructor factor accounts for approximately 34 per cent of the variability in those judgments. Furthermore, the fact that satisfaction with instructor is highly correlated with the utility scale (r = .71) offers added support for the importance of instructors in the training process. Clearly, instructors have great impact on the experience and perceptions of training participants.

A third finding of interest is that six factors emerge from the factor analysis. While the instructor factor accounts for a sizable proportion of variance, five additional factors contribute meaningful increments of variance. Also, the pattern of interrelationship among the factor derived scales supports the contention that participant reactions are multidimensional. While some correlations are high, others are low. This evidence of convergent - discriminant validity contradicts those who reduce participant reactions to unidimensional "smile sheets". It also suggests that two dimensional characterizations (e.g., affective, utility) fail to adequately capture the information available in well-developed participant reaction forms. Clearly, the constructs underlying participant reactions are more complex than those researchers contend.

Finally, examination of the scale intercorrelations yields another interesting note regarding perceptions of utility. In addition to the instructor scale, the materials and testing scales are all highly correlated with perceptions of the usefulness of training. Future research should examine the underlying components of training that contribute to participants' perceptions that it will be useful. This research suggests that the instructor, the materials and the testing practices may play an important role.

Summary

The stated objective of this research is to make progress toward an improved technology of participant reactions. Implicit in that objective is the authors' belief that there is merit in these judgments. At a minimum, the results presented here suggest that the constructs underlying participant reactions are more complex than previously suggested. It is reasonable to argue that the past failure to acknowledge the multiple dimensions of employee reactions may well account for the dismal pattern of correlation between participant reactions and other "levels" of training criteria. While the low correlation with other evaluation criteria have led some (Holman, 1996) to question the credibility of participant reactions as training criteria, the authors of this study are more optimistic about the role that participant reactions can play in systematic training evaluation. Clearly, additional work is needed to improve the understanding and measurement of participant reactions.

Future research will seek to replicate the factor structure and interrelationship of participant reactions in other data sets, and seek to further understand the constructs underlying participant reactions and their relationship to other indices of training effectiveness.

BEST COPY AVAILABLE

References

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. Personnel Psychology, 42(2), 331-342.

Alliger, G. M., Tannenbaum, S. I., Bennett, J., Winston, Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. Personnel Psychology, 50, 341-358.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. Personnel Psychology, 44(1), 1-26.

Barrick, M. R., Mount, M. K., & Strauss, J. P. (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. Journal of Applied Psychology, 78(5), 715-722.

Bassi, L.J., Benson, G., Cheney, S. (1996). The top 10 trends. Training and Development, 50, 28-42.

Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel Selection. Annual Review of Psychology, 48, 299-337.

Brayfield, A. H., & Crockett, W. H. (1955). Employee attitudes and employee performance. Psychological Bulletin, 52, 396-424.

Brinkerhoff, R.O. (1986). Achieving results from training. San Francisco: Jossey-Bass.

Fisher, C. D. (1980). On the dubious wisdom of expecting job satisfaction to correlate with performance. Academy of Management Review, 5(4), 607-611.

Hogan, J. C., Hogan, R., & Murtha, T. (1992). Validation of a personality measure of managerial performance. Special Issue: Test validity yearbook: I. Journal of Business and Psychology, 7(2), 225-237.

Holman, E. F. (1996). The flawed four-level evaluation model. Human Resource Development Quarterly, 7(1), 5-21.

Hough, L. M. (1992). The "Big Five" personality variables - Construct confusion: Description versus prediction. Human Performance, 5(1&2), 139-155.

Iaffaldano, M. T., & Muchinsky, P. M. (1985). Job satisfaction and job performance: A meta-analysis. Psychological Bulletin, 97(2), 251-273.

Kirkpatrick, D.L. (1959). Techniques for evaluating training programs. Journal of ASTD, 13, 3-9.

Kirkpatrick, D.L. (1996). Evaluating training programs: The four levels. San Francisco: Barret-Koehler.

MacKenzie, S. B., Shankster, L. J., & Kohler, S. S. (1991). Organizational citizenship behavior and objective productivity as determinants of managerial evaluations of salespersons' performance. Organizational Behavior and Human Decision Processes, 50, 123-150.

Ones, D. S., Mount, M. K., Barrick, M. R., & Hunter, J. E. (1994). Personality and job performance: A critique of the Tett, Jackson, and Rothstein (1991) meta-analysis. Personnel Psychology, 47(1), 147-156.

Phillips, J. J. (1997). Handbook of training evaluation and measurement methods. Gulf Publishing Co.

Tabachnick, B. G., & Fidell, L. S. (1996). Using Multivariate Statistics. (third ed.). Northridge, CA: HarperCollins College Publishers.

Saari, L.M., Johnson, T.R., McLaughlin, S.D., & Zimmerle, D.M. (1988). A survey of management training and evaluation practices in U.S. companies. Personnel Psychology, 41, 731-743,Tett, R. P., Jackson, D. N., Rothstein, M., & Reddon, J. R. (1994). Meta-analysis personalityob performance relations: A reply to Ones, Mount, Barrick, and Hunter (1994). Personnel Psychology, 47(1), 157-172.

Warr, P. & Bunce, D. (1995). Trainee characteristics and the outcomes of open learning. Personnel Psychology, 48, 347-376

BEST COPY AVAILABLE